Lecture 11
Parametric GLM models for over-dispersion

Today’s topics:

- Negative Binomial GLM
- Zero inflated models: ZIP, ZINB and hurdle models
- Revisit the example of the horseshoe crab dataset
- Beta-Binomial GLM

Over-dispersion in the Poisson model

- Poisson regression assume that $\mathrm{Var}[y_i|X_i] = \mathbb{E}[y_i|X_i]$
- Over-dispersion: in practice, the counts y_i can be noisier than assumed in the Poisson distribution

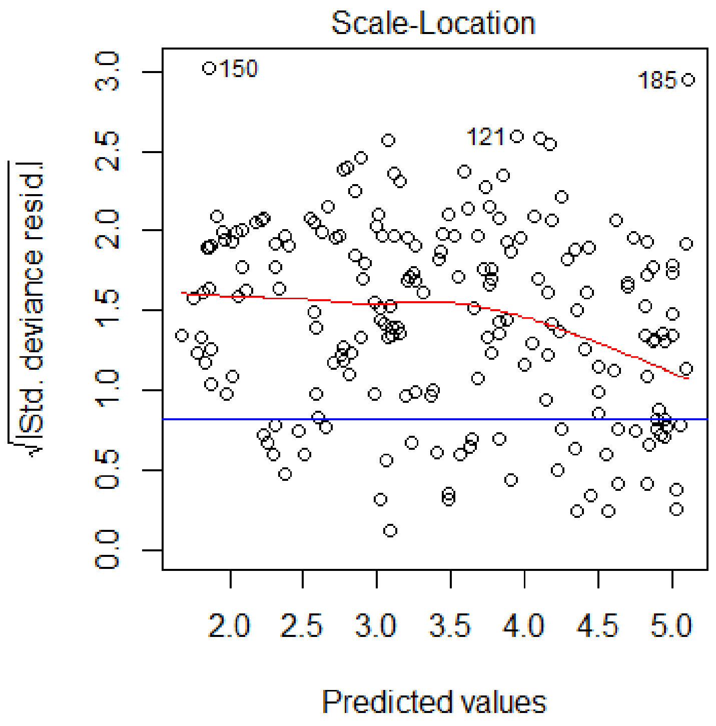
- For instance, if $\log(\lambda_i) = X_i^T \beta + \epsilon_i$ indicating that X_i can not fully explain λ_i. Then

$$E(y_i) = E[E(y_i \mid \lambda_i)] = E(\lambda_i)$$

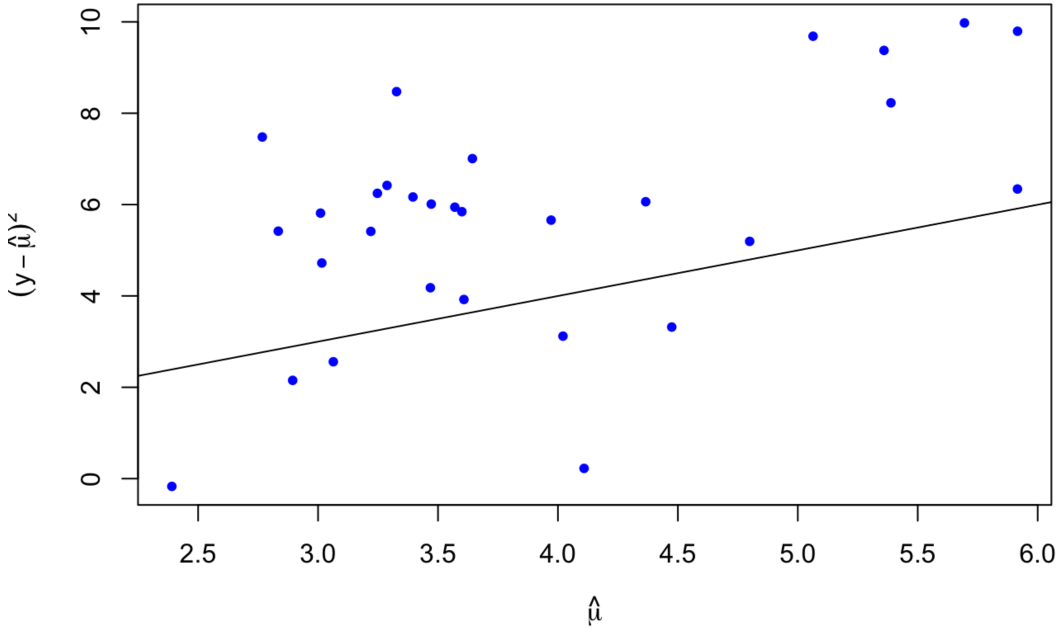
while

$$\mathrm{Var}(y_i) = E[\mathrm{Var}(y_i \mid \lambda_i)] + \mathrm{Var}[E(y_i \mid \lambda_i)] = E(\lambda_i) + \mathrm{Var}(\lambda_i) > E(y_i)$$

Over-dispersion examples



https://stats.stackexchange.com/questions/331086/investigate-overdispersion-in-a-plot-for-a-poisson-regression



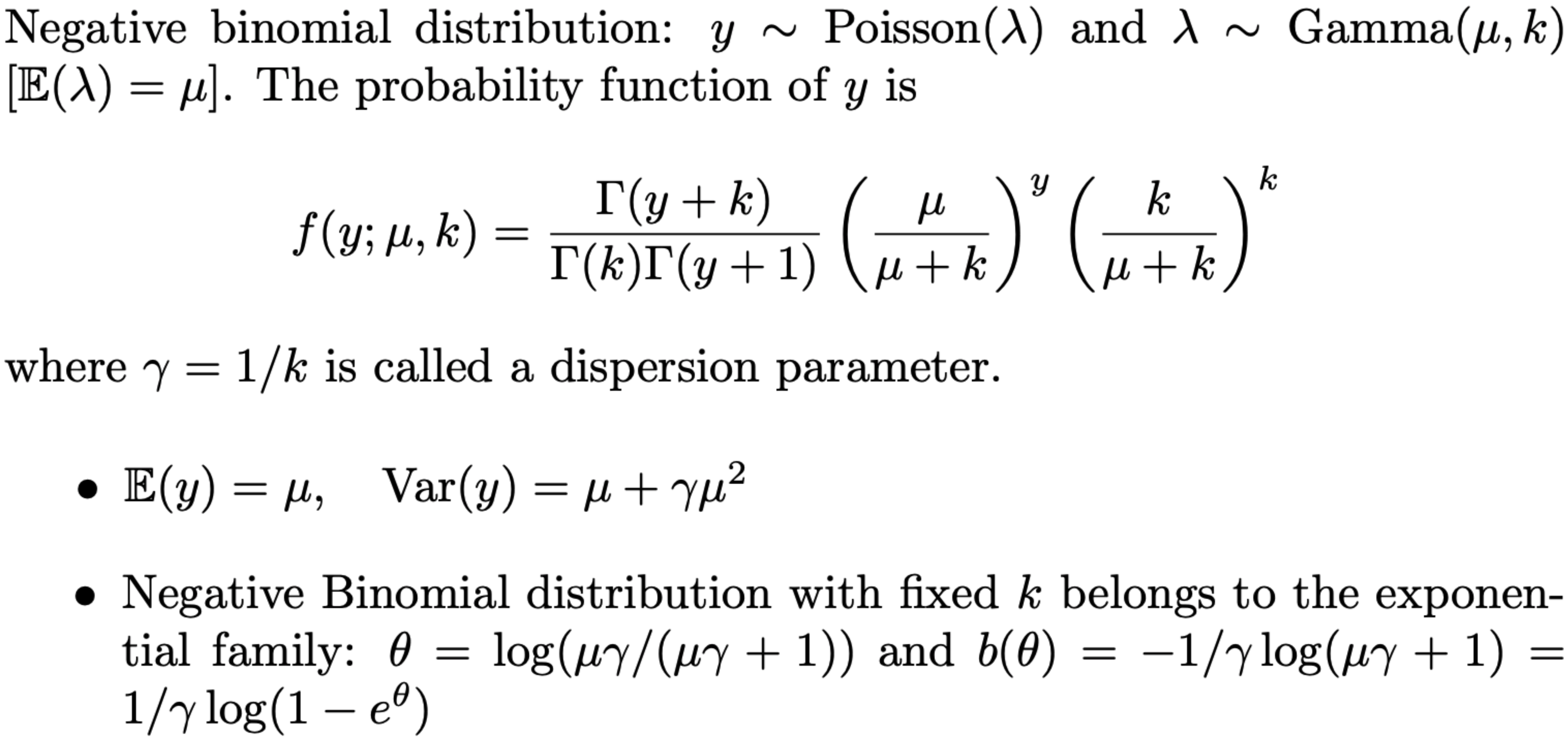
https://towardsdatascience.com/adjust-for-overdispersion-in-poisson-regression-4b1f52baa2f1

Over-dispersion in the Poisson model

- For example, we saw the over-dispersion issue in the horseshoe satellites dataset in Data Example 1 and homework 1, 1.22(a).
- Over-dispersion happens in Poisson and Binomial (Multinomial) GLM models as the variance is completely determined by the mean.
- There is no over-dispersion issue in linear models as linear models has an extra dispersion parameter.

- We will talk about semi-parametric solutions for over-dispersion issues in next lecture

Negative binomial distribution



Negative binomial distribution: $y \sim \text{Poisson}(\lambda)$ and $\lambda \sim \text{Gamma}(\mu, k)$ $[\mathbb{E}(\lambda) = \mu]$. The probability function of y is

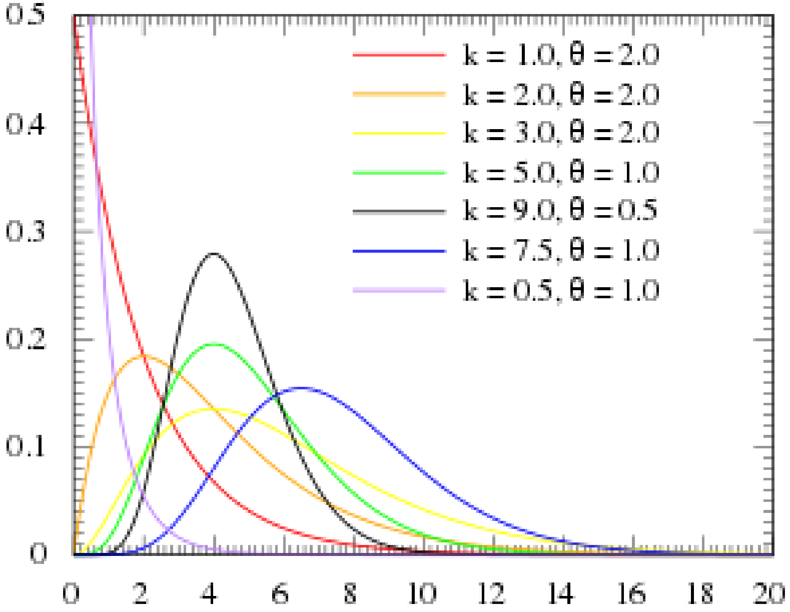
$$f(y; \mu, k) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \left(\frac{\mu}{\mu+k}\right)^{y} \left(\frac{k}{\mu+k}\right)^{k}$$

where $\gamma = 1/k$ is called a dispersion parameter.

- $\mathbb{E}(y) = \mu, \quad \text{Var}(y) = \mu + \gamma\mu^2$
- Negative Binomial distribution with fixed k belongs to the exponential family: $\theta = \log(\mu\gamma/(\mu\gamma + 1))$ and $b(\theta) = -1/\gamma \log(\mu\gamma + 1) = 1/\gamma \log(1 - e^{\theta})$

Negative binomial distribution

- It is defined as compound distribution (Gamma-Poisson mixture)



- Mean and variance of a Gamma distribution:

$$\mu = k\theta, \qquad \mathrm{Var}(\lambda) = k\theta^2 = \frac{\mu^2}{k} = \gamma\mu^2$$

- For NB distribution

$$\mathbb{E}(y) = \mu, \quad \mathrm{Var}(y) = \mu + \gamma\mu^2$$

Negative binomial GLM

- We assume that

  $$y_i \sim \mathrm{NB}(\mu_i, k_i)$$

  with the link function $g(\mu_i) = X_i^T \beta$.

  - Typically, we assume that all samples share the same dispersion, so $\gamma_i = \frac{1}{k_i} = \gamma$.
  - As an extension of the Poisson GLM, a common link for NB GLM is still the loglinear link: $g(\mu_i) = \log(\mu_i)$
  - Score equation for β

$$\sum_i \frac{y_i - \mu_i}{\mu_i + \gamma\mu_i^2} \mu_i x_{ij} = \sum_i \frac{y_i - \mu_i}{1 + \gamma\mu_i} x_{ij} = 0$$

Negative binomial GLM

A bit about the inference:

- The hessian matrix has the term

$$\frac{\partial^2 L(\boldsymbol{\beta}, \gamma; \boldsymbol{y})}{\partial \beta_j \partial \gamma} = -\sum_i \frac{(y_i - \mu_i)x_{ij}}{(1 + \gamma\mu_i)^2} \left(\frac{\partial \mu_i}{\partial \eta_i}\right).$$

Thus, $E(\partial^2 L/\partial \beta_i \partial \gamma) = 0$ for each j, and $\boldsymbol{\beta}$ and γ are orthogonal parameters

- the asymptotic variance of $\hat{\beta}$ would be the same no matter γ is estimated or known (Agresti book chapter 7.3.3)

$$\widehat{\mathrm{Var}}(\hat{\beta}) = (X^T \hat{W} X)^{-1}$$

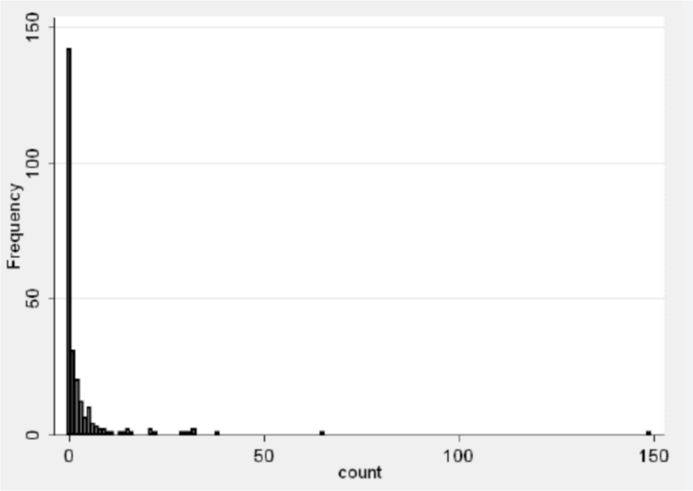
  - $w_i = \mu_i/(1 + \gamma\mu_i)$

Zero-inflated counts

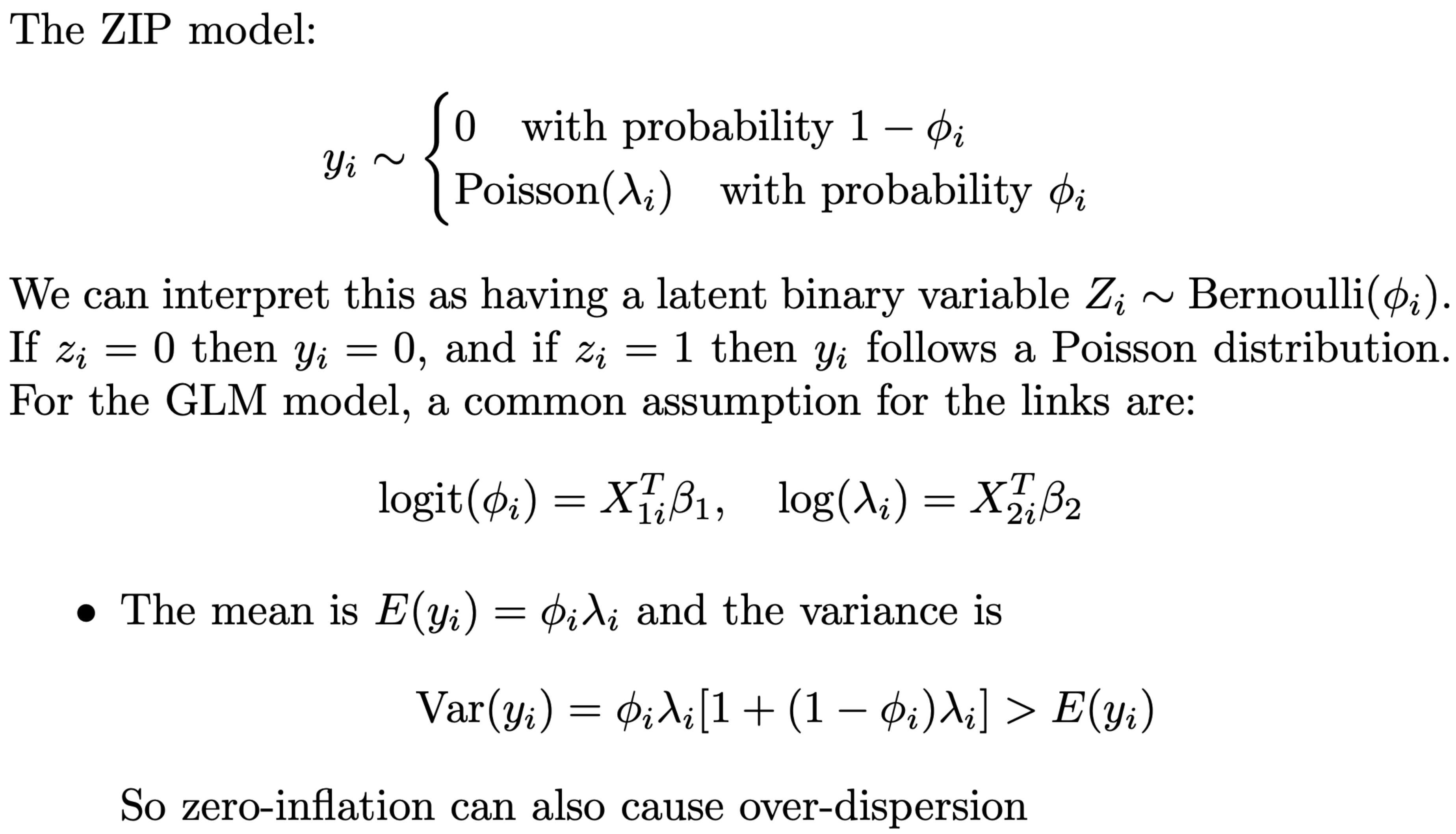
For a Poisson distribution $y \sim \text{Poisson}(\mu)$: $P(y=0) = e^{-\mu}$

For a Negative Binomial distribution $y \sim \text{NB}(\mu, k)$: $P(y=0) = \left(\frac{k}{\mu+k}\right)^k$

- In practice, there may be way more 0 counts than what these distributions can allow
- Example: y_i is the number of times going to a gym for the past week and there may be a substantial proportion who never exercise



Zero-inflated Poisson models



The ZIP model:

$$y_i \sim \begin{cases} 0 & \text{with probability } 1-\phi_i \\ \text{Poisson}(\lambda_i) & \text{with probability } \phi_i \end{cases}$$

We can interpret this as having a latent binary variable $Z_i \sim \text{Bernoulli}(\phi_i)$. If $z_i = 0$ then $y_i = 0$, and if $z_i = 1$ then y_i follows a Poisson distribution. For the GLM model, a common assumption for the links are:

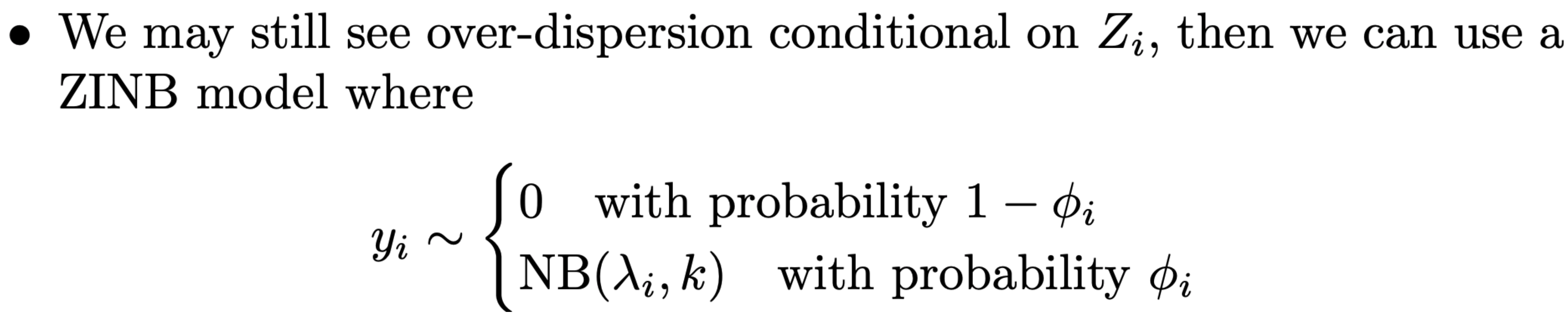
$$\text{logit}(\phi_i) = X_{1i}^T\beta_1, \quad \log(\lambda_i) = X_{2i}^T\beta_2$$

- The mean is $E(y_i) = \phi_i\lambda_i$ and the variance is

$$\text{Var}(y_i) = \phi_i\lambda_i[1 + (1-\phi_i)\lambda_i] > E(y_i)$$

So zero-inflation can also cause over-dispersion

Zero-inflated Negative Binomial models



- We may still see over-dispersion conditional on Z_i, then we can use a ZINB model where

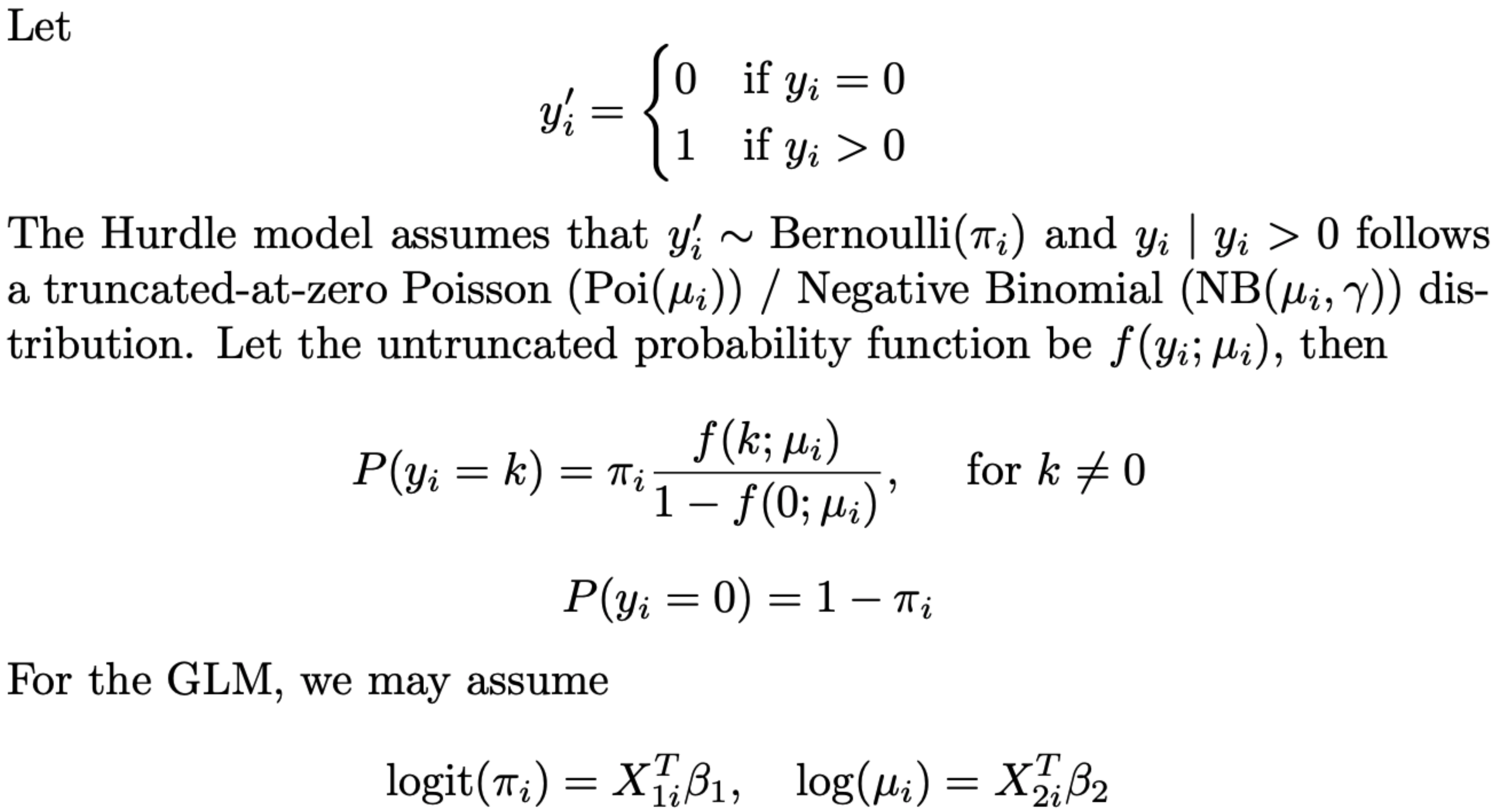
$$y_i \sim \begin{cases} 0 \quad \text{with probability } 1 - \phi_i \\ \text{NB}(\lambda_i, k) \quad \text{with probability } \phi_i \end{cases}$$

- We can still use MLE to solve both the ZIP and ZINB model

- The ZIP/ZINB model do not allow zero deflation.

The Hurdle model

- The Hurdle model separates the analysis of zero counts and positive counts.



Let

$$y_i' = \begin{cases} 0 & \text{if } y_i = 0 \\ 1 & \text{if } y_i > 0 \end{cases}$$

The Hurdle model assumes that $y_i' \sim \text{Bernoulli}(\pi_i)$ and $y_i \mid y_i > 0$ follows a truncated-at-zero Poisson ($\text{Poi}(\mu_i)$) / Negative Binomial ($\text{NB}(\mu_i, \gamma)$) distribution. Let the untruncated probability function be $f(y_i; \mu_i)$, then

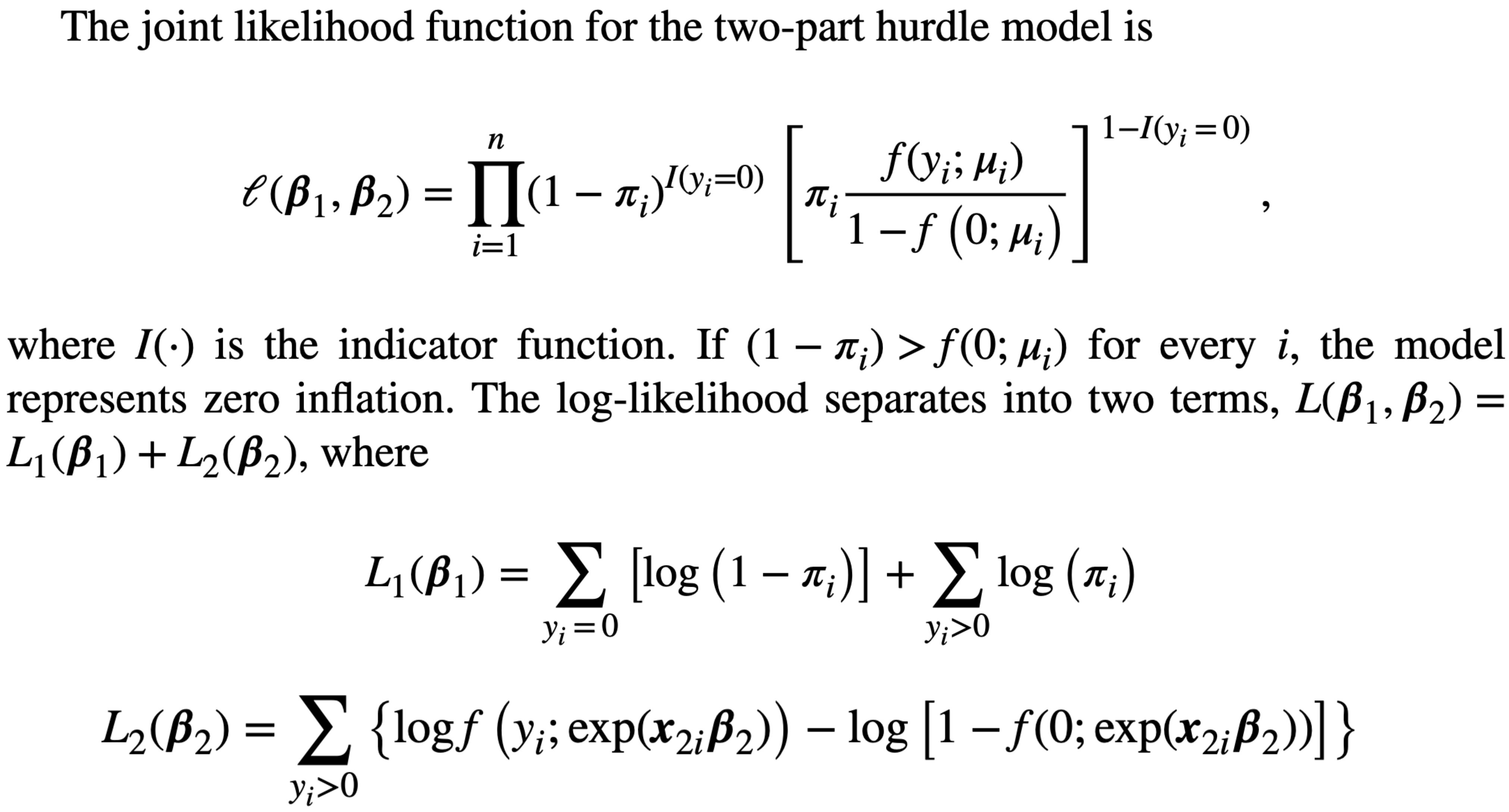
$$P(y_i = k) = \pi_i \frac{f(k; \mu_i)}{1 - f(0; \mu_i)}, \quad \text{for } k \neq 0$$

$$P(y_i = 0) = 1 - \pi_i$$

For the GLM, we may assume

$$\text{logit}(\pi_i) = X_{1i}^T \beta_1, \quad \log(\mu_i) = X_{2i}^T \beta_2$$

The Hurdle model



The joint likelihood function for the two-part hurdle model is

$$\ell(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \prod_{i=1}^{n}(1-\pi_i)^{I(y_i=0)} \left[\pi_i \frac{f(y_i;\mu_i)}{1-f\left(0;\mu_i\right)}\right]^{1-I(y_i=0)},$$

where $I(\cdot)$ is the indicator function. If $(1-\pi_i) > f(0;\mu_i)$ for every i, the model represents zero inflation. The log-likelihood separates into two terms, $L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = L_1(\boldsymbol{\beta}_1) + L_2(\boldsymbol{\beta}_2)$, where

$$L_1(\boldsymbol{\beta}_1) = \sum_{y_i=0}\left[\log\left(1-\pi_i\right)\right] + \sum_{y_i>0}\log\left(\pi_i\right)$$

$$L_2(\boldsymbol{\beta}_2) = \sum_{y_i>0}\left\{\log f\left(y_i;\exp(\boldsymbol{x}_{2i}\boldsymbol{\beta}_2)\right) - \log\left[1-f(0;\exp(\boldsymbol{x}_{2i}\boldsymbol{\beta}_2))\right]\right\}$$

Revisit the horseshoe crab data

- Check Example6 R notebook

Violation of the variance assumptions in GLM

In earlier models, we typically have assumptions on the variance of $y_i|X_i$

- Gaussian linear model: $\mathrm{Var}(y_i) = \sigma^2$
- GLM with Binomial / Multinomial / Poisson models: fixed mean-variance relationship

As we saw earlier, real data can have over-dispersion / under-dispersion or unequal variances, which violates these variance assumptions

- With wrong variance assumption but correct mean assumption (link function)
  - Typically still get consistent point estimate $\hat{\beta}$
  - Inference on $\hat{\beta}$ can be heavily impacted

Variance inflation in binomial GLM

For the ungrouped Binary data, previous Binary GLM assumed that conditional on having the same X_i, the y_i are i.i.d. Bernoulli trials.

What if the samples within each group are correlated?

- Analogous to the Poisson case, we can have the scenario

$$y_i \sim \text{Binomial}(n_i, p_i) \text{ but } \text{logit}(p_i) = X_i^T \beta + \epsilon_i$$

- Such a hierarchical model leads to variance inflation:

$$\text{Var}(y_i) > n_i p_i (1 - p_i)$$

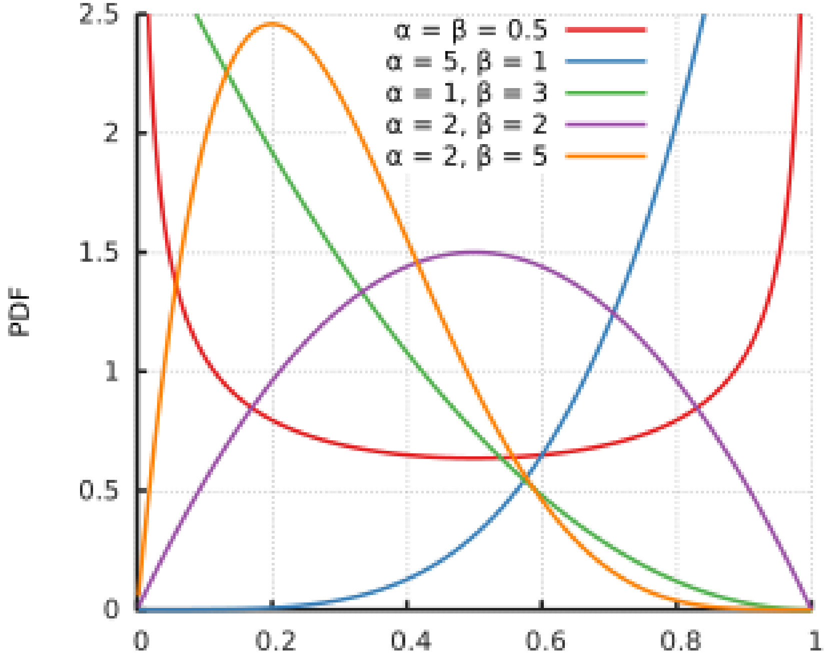
- If you treat y_i as a sum of Bernoulli variables $y_i = \sum_j Z_{ij}$ where $Z_{ij} \sim \text{Bernoulli}(p_i)$, then randomness in p_i causes dependence among Z_{ij}.

Beta-binomial distribution

- The Beta-binomial distribution assumes that $y \sim \text{Binomial}(n, p)$ and $p \sim \text{beta}(\alpha, \beta)$. The beta distribution of p has the density function:

$$f(p; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

- Beta distribution



- Mean and variance of a Beta distribution:

$$\mu = \frac{\alpha}{\alpha+\beta},$$

$$\text{Var}(p) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \mu(1-\mu)$$

- For Beta-binomial distribution distribution

$$E(y) = n\mu, \quad \text{Var}(y) = n\mu(1-\mu)\left[1+(n-1)\rho\right]$$

where $\rho = 1/(\alpha+\beta+1)$.

Beta-binomial GLM

- We assume that

$$y_i \sim \text{Beta-binomial}(n_i, \mu_i, \rho)$$

with the link function $g(\mu_i) = X_i^T \beta$. $\mathbb{E}(y_i) = n_i \mu_i$

- As before, we assume that all samples share the same dispersion, so there is only one unknown dispersion parameter ρ.
- A common link for Beta-binomial GLM is still the logit link:

$$\text{logit}(\mu_i) = X_i^T \beta$$

- Both β and ρ are unknown but we can estimate using MLE.